基于主题河的网络舆情可视化关联分析方法*

詹 建 高民权 （兰州大学信息科学与工程学院 甘肃 730000）

摘要 文章介绍了一种以主题河（ThemeRiver）为核心的网络舆情可视化关联分析方法，探讨了该模型的构建策略，原型系统的功能设置、操作方法、设计与实现。实证结果表明，该方法不但能从整体上展现舆情在时间上的类别变化关系，而且能从多种角度显示类属、关系、热点话题等细节，能够很好地帮助用户发现舆情变化背后的原因和规律。

关键词 主题河 可视化 网络舆情

ThemeRiver – based Internet Public Opinion Visualization Correlation Analysis Methods

Zhan Jian Gao Minquan （School of Information Science and Engineering of Lanzhou University, Gansu, 730000）

Abstract This paper introduces a visualization model that oriented Internet public opinion analysis based on ThemeRiver. The authors specifically discuss constructing strategy of the visualization model, functional module and operation method of prototype system, design and implementation. The empirical results show that this method not only show the changes relationship between public opinion in time from the overall category, but display the details such as generic, relationships, hot topics and so on from a variety of angles that can help users find a good reason behind the change in public opinion and laws.

Keywords ThemeRiver, visualization, Internet public opinion

1 引言

网络舆情是由于各种事件的刺激而产生，通过互联网传播的人们对于该事件的所有认知、态度、情感和行为倾向的集合[1]，其实质是大众公共观点在互联网媒介中的反映。

2014年1月16日，中国互联网络信息中心（CNNIC）发布《第33次中国互联网络发展状况统计报告》，数据显示，截至2013年12月，中国网民规模已达6.18亿，互联网普及率达到45.8%[2]。中国网民数量已经稳居世界第一。在此背景下，基于各种Web2.0技术的社会媒体产生的网络舆情，不仅影响着社会群体的观点，也深深影响着整个社会的行为。探索各种类型网络舆情的特征和演变机制，对于增强新形势下社会公共治理能力和商业竞争策略具有重要意义。但是，通常人们难以从海量的网络文本中直接获取到有效的信息，因此大大限制了分析信息的能力。信息可视化是一种通过利用人类的视觉能力，来理解和分析抽象信息的意义，从而加强人类的认知能力的途径[3]。

本文提出一种网络舆情可视化分析的方法，并设计了相应的原型系统。该系统能够以一张图为核心对原始数据进行多角度数据展示，用户可以探索式交互地了解一段时间内网络某事件中舆情随时间的演化过程、演化的细节、热点话题以及分析推动种种变化的原因。本文以兰州2014年发生的一起公共事件为例，对本方法的可用性及其有效性进行了验证。

2 研究回顾

网络舆情作为数据表现为基于时间序列的具有社会属性的海量短文本集合。舆情分析一大核心目标

★本文系国家自然科学基金项目“公众参与的社区减灾管理：公共知识的形成机理及其人—机协作系统研究”（编号：71373108）及中央高校基本科研业务费专项资金项目“基于社会计算技术的社区减灾协商研究”（编号：14LZUJBWZD012）的研究成果之一。

专题研究

是从这些纷繁复杂的文本中抽取涉及的主题。对文本主题在时间上的建模可以简单地分为两大类：一类是将时间视为连续随机变量建模，另一类是基于离散化的时间点构建动态贝叶斯网络模型[4]。

国内计算机和信息科学学者在文本可视化方面开展了一定的基础方法以及应用研究。清华大学孙茂松指出可视化方法是文本挖掘的重要组成部分，文本可视化综合了文本分析、数据挖掘、数据可视化、数据集图形学、人机交互、认知科学等学科的理论和方法，为人们提供了一种理解海量复杂文本的内容、结构和内在规律等信息的有效手段[5]。在文本可视化方面，武汉大学的周宁团队进行了大量的研究工作，对文本可视化技术进行了综述，细致探究了图符标识法、高维空间描述法[6-7]，提出文本信息可视化通用模型[8]，给出了基于非线性映射的可视化文本聚类的方法[9]。张兆锋、陈颖、安海忠、刘永等学者对文本挖掘中的信息结果和关系可视化进行了多角度的应用阐释[10-13]。

目前，我国针对网络舆情的研究非常活跃，积累了大量的成果。但是，结合舆情自身特点的可视化分析研究工作并不多。部分学者开展了一些应用层次上的研究，比如武汉大学申莹对舆情中的话题聚类和一般可视化呈现做了探索和实现[14]。信息工程大学郭建忠等人研究了舆情在GIS上的反映[15]。公安大学许星等人利用斯坦福大学开发的protovis对微博中舆情信息可视化进行了技术可行性方面的初步尝试[16]。总的来说，系统化、基础性的研究工作还比较缺乏。

传统的文本分析办法难以反映舆情文本之间的时间属性。主题河(ThemeRiver)是一种被证明为有效的反映文本之间的时间属性的方法[17]。在这种可视化方法中，时间被表示为从左往右的一条水平轴，然后用不同的颜色条带代表不同的主题，条带的宽度代表该主题在该时间的一个度量（例如主题的提及频率），使用这种表示方式的最大优点是人们可以很容易地跟踪任何一个主题在量上随时间的变化。此外，也能容易地比较不同的主题在同一个时刻的相对规模大小。经典主题河技术发展主要侧重于集成更多维的数据，而较少涉足关联分析研究。

3 基于主题河的网络舆情可视化关联分析模型

3.1 网络舆情事件的数据模型

本文假定文本集合包含的每一条发言都具有一定的指向性和倾向性，参考文献[18-19]对舆情事件的数据模型定义如下：该模型命名为PO，$PO=(D, P, J, W)$。

定义1 给定一个待分析文本集合D，$d_i\in D$，D中的成员d_i表示一条该事件相关发言，$d_i=\{id, ctime, j, c, t_0, t_a\}$。其中，$id$表示发言编号，是该发言在文本集合中的唯一标识；$ctime$为发布时间；$J$表示发言人集合；$C$表示发言内容特征表达集合；$t_0$表示发言中观点所指向的目标对象；$t_a$表示发言类型，有原创、评论、回复3种，分别取值为0、1、2。

C常用向量空间模型表示方法计算求得，t_0需要用自然语言实体抽取技术获取。

定义2 舆情事件基本概要$P=\{Rank, Value_r, Value_s, S\}$，其中，$Rank$表示该舆情事件的热度；$Value_r$表示事件相关民众利益类属；$Value_s$表示事件相关民众诉求类属；$S$代表事件发展所处的状态。

$Rank$评估算法一般认为热度是主题的出现频率和爆发性的函数；$Value_s$通过文本集合分类技术获取；S取值为认知、态度表达、行动的概率分布[18]。

定义3 J为发言人集合，j_i为发言者基本个人信息(Profile)，包括性别、年龄、地域、兴趣、教育程度、工作等。

定义4 舆情社会性行为$W=\langle J, Y, E\rangle$，其中，Y是主题集合；E代表d_i间蕴含的交互关系集合。$e=\{d_i, d_j, t\}$，设d_i，d_j为具有相关性关系的发言内容，若d_i按时间较d_j早出现，则d_i为d_j的前项。t是前后项所表征的发言持有人之间产生的交互关系的映射，$t=\{g: j_i\rightarrow j_j\}$，$j_i$，$j_j$为发言人。

3.2 网络舆情事件分析的任务模型

网络舆情分析的技术目标主要为：抽取文本集合D中所涉及的对象特征，对特征归并，对观点X进行分类。分析结果得到一个四元组(J, O, X, M)，J是观点持有者，O指舆情事件实体，实体O可以是产品、政策、机构、人、组织、事件等。数据分析最终期望发现D中的所有隐含信息[20]。

定义1 基于特征的观点挖掘：设对象D为一个有限特征集合，$W=\{w_1, w_2, \cdots, w_n\}$表示特征项空间。观点$X=\{x_1, x_2, \cdots, x_n\}$，其中$x_i\in F$，每一个$f_i\in F$可以表示为一个词的有限特征集合在$W$上的投影，二者等价。对于观点持有人$J_i$的一条发言，可认为$J_i$选取一个$U_k$描述一个对象$O$的属性，同时表达了一个观点$X_k$。

定义2 观点的结构M：观点的结构表示为(T, V, D)，T是一个层次型实体的子集合或者分类，每一个实体有自己的成员及其属性，V是实体T的属性集合，也可以用特征来表示T。$d_i\in D$，定义为一系列特征词集合$d_i=(s_1, s_2, \cdots, s_M)$。

3.3 网络舆情事件分析的可视化关联分析模型

基于以上舆情事件的数据模型和任务模型，本节给出舆情事件的可视化关联分析模型，使每一个具体舆情事件分析动作可用该模型的一个程序运行期实例表示。

定义1 可视化操作F

舆情事件可视化变换可分为基本变换和复合变换两种类型。基本变换由分析员对关联信息图的操作产生，复合变换由程序经基本变换关联操作后自动产生。“变换”实体的属性包括(aId, aNum, aType)：aId为

唯一性识别标识符；aNum 表示所关联事件的个数；aType 表示关联方式，1 表示时间关联，2 表示发言者身份关联，3 表示主题关联。

$F = F_1 \cup F_2 \cup F_3$，F 表示针对关联主体的操作及主客体变换之间的关系。

F_1 = {动态过滤、概览明细、热点上下文、视图关联、投影变形}。F_1 定义可视化分析时的五个基本交互式操作。

$F_2 = \{g: F_1 \rightarrow c_{sa}\}$，$F_2$ 定义操作序关系，g 表示关联主体执行一个关联动作必须满足的条件，c_{sa} 表示关联客体可能的行为。

$Q = \{q: F_3 \rightarrow \Delta_\gamma\}$，$F_3$ 定义序变化函数，表示动作之间的时间序列上的关系，Δ_γ 表示操作动作之间的时间差。

定义 2 可视化映射规则 R

映射规则完成舆情事件实体属性及操作的图形化表达。基本属性映射如表 1，基本操作映射参考 F_1。

表 1 实体基本属性可视化映射

事件实体属性	图形化表达
关系	图节点
层次	树节点
序列	一维坐标
属性值	比例图或者多维坐标
频率	图形尺寸

定义 3 初始图形布局 S_0

S_0 对应于一个舆情事件的初始状态，设定关联体、关联客体、关联函数、关联主体值 $S_{ss}(0)$，则生成关联客体相应状态 $F\{S_0(0)\}$，然后通过上述映射规则完成图形布局。

4 可视化关联分析设计与实现

4.1 可视化关联分析功能结构设计

“主题河”体现了时间轴上不同类属的量的比例变化关系，是一种非常合适的舆情数据可视化分析工具。“主题河”不仅能够使用户了解某一个主题的演化进程，还能够对数据集合整体态势有较为直观的认识，有助于帮助人们分析该主题的演化、发展情况和近期关注热点。

我们以“主题河”为基础进行了关联扩展，提出了一种新的网络舆情可视化分析方法。该方法核心视觉元素为主题河、扫描线和关联功能图。通过扫描“主题河”驱动同一时间点的关联功能图数据(见图 1)。

本可视化分析方法定义四种基本类型的关联功能图：

(1)明细表格：查看明细数据；

(2)网络图：观察数据之间的各种关系属性(传播，核心节点，用户间关系等)；

(3)标签云：概览数据全貌，热点话题等；

(4)分类图：多值分类，对观察对象进行进一步分类聚类。文本数据经常需要和其他相关的非文本数据融合在一起进行关联分析。例如微博数据既包含了文本数据，也包含了用户的一般性资料，如地理信息、年龄段、性别等结构性的非文本数据，对这些信息的分类常常可以用来解释不同舆情形成现象。

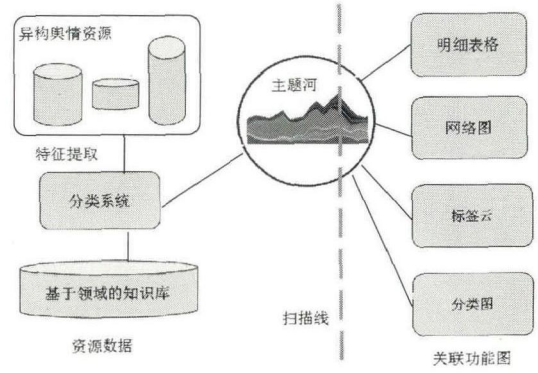

图 1 可视化关联分析设计功能结构图

4.2 可视化关联分析交互流程设计

借鉴可视化工具 Prefsue 的思想[21]，设计可视化关联分析交互的基本流程如下(见图 2)：

(1)载入数据。将待处理文档集合以可视化映射规则 R 中属性映射为指导转换成绘图引擎的内嵌图形化结构数据，建立待处理数据对象模型实例。

(2)建立可视化对象。可视化对象负责将装载的数据映射为可视化元素，根据关联分析模型所定义的关联主客体要素建立数据动态响应关系。

(3)可视化服务响应实例生成和实例注册。通过前述操作自动触发实例映射，生成实例，在系统中注

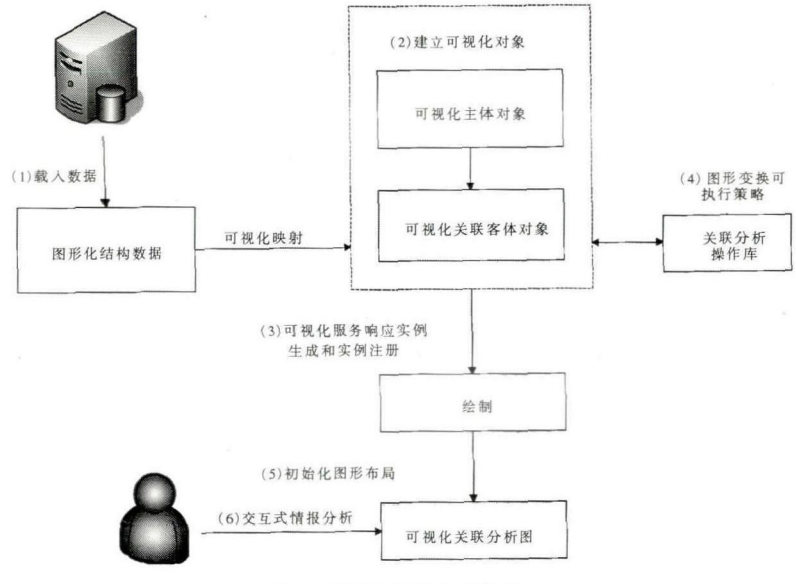

图 2 可视化交互基本流程

册实例，服务响应实例负责绘制可视化元素。

(4)建立图形变换可执行策略。以可视化映射规则 R 中属性映射和操作映射为指导提供操作的规范。本步骤根据舆情信息的类型、频率和语义关系等属性信息分别赋予不同图形特征，及其特征变换方法，比如可视化元素的位置、形状、大小、颜色和投影的设置。本步骤生成关联分析操作库，其执行逻辑如下(见图 3)。

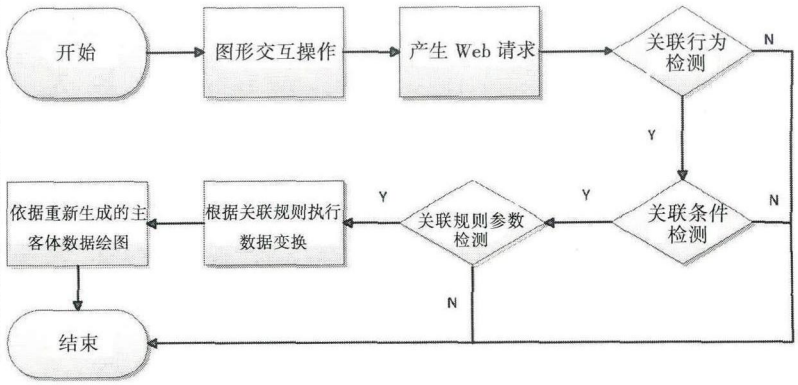

图 3 图形变换策略执行流程

(5)初始化显示。由 S_0 形成初始图形布局。

(6)交互式情报分析。完成舆情信息探索过程。

4.3 系统实现技术框架

前端系统整体界面设计选用 Bootstrap，Bootstrap 是 Twitter 推出用于前端开发的开源工具包，它是一个基于 Jquery 的 CSS/HTML 框架。数据可视化部分选用 D3，D3 是近几年出现的具有广泛成功案例的可视化 JS 库，它被很多其他的 Web 系统所使用，它允许绑定任意数据到 DOM，然后将数据驱动转换应用到 Document 中。

系统数据处理后端功能模型由如五部分组成，即数据采集、数据清洗、数据存储、数据分析和数据可视化。后端功能实现选用基于 Python 的技术栈。

文本处理方法方面，使用了 GitHub 上开源的 JIEBA 分词系统，该系统基于 Python，支持三种分词模式。本文中的研究重点不在于具体的统计学习算法，因此直接集成现成的开源算法包中需要的算法，用到的工具包括：NetworkX(图模型)、scikit - learn(机器学习)、Pandas(统计处理)、Pytables(时间序列处理)。

4.4 系统使用基本流程

系统使用基本流程包括七个步骤：(1)领域问题设定；(2)选取数据源；(3)采集数据；(4)文本处理；(5)根据可视化目标模型设置相应的配置文件；(6)基于主题河，配合合适的关联功能图选取，利用扫描线进行数据关联分析；(7)分析结果，发现数据中蕴含的问题特征、效应、机制等深层知识。

5 兰州自来水苯超标事件实证分析

5.1 事件背景

兰州自来水苯超标事件指的是兰州市威立雅水务集团公司出厂水及自流沟水样中苯含量严重超标。2014 年 4 月 11 日，据威立雅水务集团公司检测显示，4 月 10 日 17 时出厂水苯含量高达 118 微克/升，4 月 10 日 22 时自流沟(自来水一分厂与二分厂之间中间段)苯含量为 170 微克/升，4 月 11 日凌晨 2 时检测值为 200 微克/升，均远超出国家限值的 10 微克/升。4 月 11 日 11 时，兰州市已停运北线自流沟，排空受到污染的自来水；南线输水管道正常供水；兰州官方特别提示，自来水不宜饮用，其他生活用水不受影响。4 月 12 日，根据调查，造成兰州自来水苯超标系中国石油天然气公司兰州石化分公司一条管道发生原油泄漏、污染了供水企业的自流沟所致。4 月 14 日，兰州四区全部解除应急措施，全市自来水恢复正常供水[22]。

5.2 数据来源

于 2014 年 4 月 26 日，对百度贴吧用关键字“兰州苯”进行搜索，然后用爬虫抓取了相关发言及其跟帖共计 5328 条，将数据存储为 JSON 格式。

5.3 数据预处理

本阶段为后续工作做准备：(1)清除 JSON 文件中不必要的数据；(2)提取文件内容：ID、发言时间、发言内容；(3)对发言内容进行分词；(4)利用大连理工大学研制的“情感词典库”做特征项过滤；(5)标注训练数据。

5.4 利用贝叶斯方法进行情感分类

朴素贝叶斯分类算法是以贝叶斯理论为基础的一种在已知先验概率与条件概率情况下得到后验概率的文本分类方法，其分类算法实现比较简单，分类效率也比较高，在文本分类方面表现比较好。具体算法如下：

当分类文本用 N 维特征向量 $d = d(w_1, w_2, \cdots, w_X)$ 表示，设 k 个训练样本集 $(C_1, C_2, \cdots, C_k)$，每个类别 Ci 的先验概率为 $P(C_i)$，则

$$P(C_i) = \frac{C_i \text{ 类的训练样本数}}{\text{训练样本集总数}}$$

5.5 其他处理

(1)层次聚类数据生成。用 K 均值聚类方法对不同时间段的发言分类，存入数据库。

(2)网络图关系数据生成。网民之间发言的相互关系以矩阵存储太过稀疏，因此将相关性关系用 JSON 格式以链表形式存入数据表的长文本字段。

(3)标签云数据生成。用互信息方法进行词频统计后的特征词选取,计算结果以文本文件形式保存。

5.6 可视化分析

在系统配置文件中将分类项(关联主体)设置为上述处理结果数据库中的情感分类表。运行程序,可以看到网民情感随着时间的变化趋势(见图 4)。

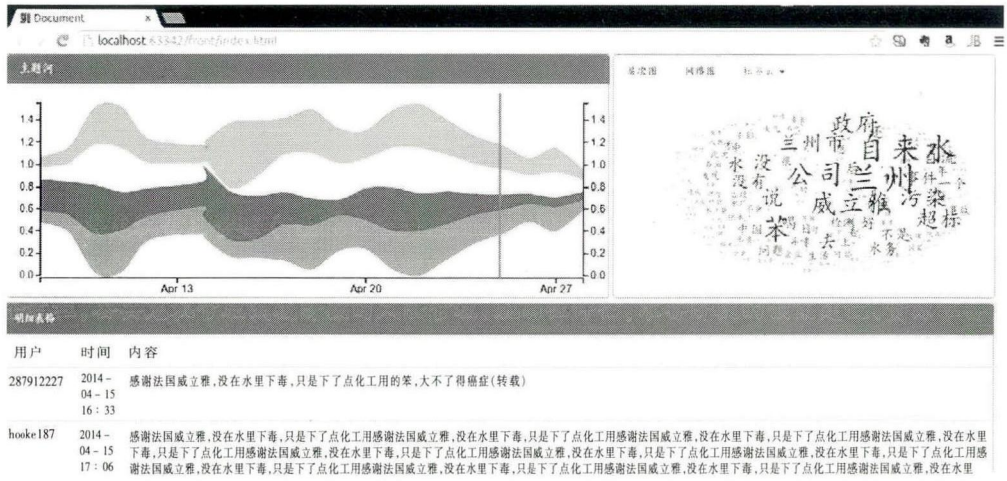

图 4 从上到下情感类别依次为“批评”、“冷静”、“恐惧”、“愤怒”

选取“标签云”功能,移动扫描线(图 4 中坐标区间竖线)到需要观察的时间点,可以看到网民在该时间段对此事情议论的热点话题(图 4 右边部分)。

在系统配置文件中将“网络分析项”(关联客体)设置为“意见领袖”,选取“网络分析”功能,移动扫描线到需要观察的时间点,从网络图中可以较清楚看到该时间段内主导、带动讨论的核心节点,也就是意见领袖(见图 5 右侧)。

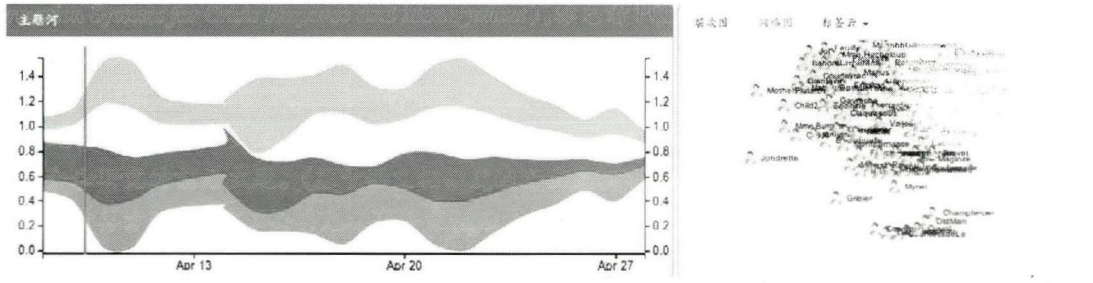

图 5 意见领袖

在系统配置文件中将“网络分析项”设置为“观点聚类”,选取“网络分析”功能,移动扫描线到需要观察的时间点,从层次图中看到该时间段讨论内容的聚类结果。结果显示,聚类较理想的将网民发言分成了四个具有明显区分度的不同类别:情绪发泄、事实陈述、质问政府、其它(见图 6)。

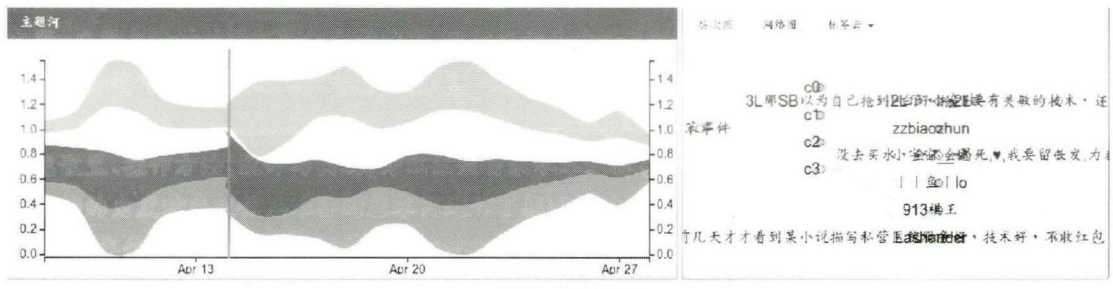

图 6 观点聚类

结果表明,该可视化方法能较好地反映现实情况。例如,在文本流比较明显的两次情绪反弹的转折点中,均是政府相关单位的不当言论使得网民对该事件的关注度再次增大,从而使得该话题的活跃度明显增加。且进一步观察可以看出,互联网用户的情感从刚开始的由担心健康安全产生的恐惧心理比例较大,渐渐地变为对相关部门的愤怒主导,这也比较符合人们的认知常识。

6 结语

本文介绍了网络舆情分析的背景和现实需求,引入了一种可视化分析的思路。然后,基于前述方法进行了一个完整实验,从而为我们方法的可行性和有效性进行了验证,也提供了实践方面的依据。文章所涉

及的工作还有几方面需要进一步研究，如可视化模型中更为丰富的图形映射集、关联规则中的因果函数的有效获取、数据接口设计的完备性等。

参考文献

[1] 曾润喜.网络舆情信息资源共享研究[J].情报杂志，2009(8):187-191.

[2] 中国互联网信息中心.第33次中国互联网络发展状况统计报告[EB/OL].[2014-06-15] http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/.

[3] Wolfe J M, Friedmanhill S R, Bilsky A B. Parallel-processing of part-whole information in visual-search tasks[J]. Perception & Psychophysics, 1994,55(5): 537-550.

[4] Wolfgang Müller, Heidrun Schumann. Visualization methods for time-dependent data - An overview[C]. Proceedings of the 2003 Winter Simulation Conference. USA: WSC, 2003: 737-745.

[5] 唐家渝,刘知远,孙茂松,等.文本可视化研究综述[J].计算机辅助设计与图形学学报,2013,25(3):273-285.

[6] 周 宁,张玉峰,张李义.信息可视化与知识检索[M].北京:科学出版社,2005:1-3.

[7] 周 宁,程红莉,吴佳鑫.信息可视化的发展趋势研究[J].图书情报工作,2008,52(8):35-38.

[8] 周 宁,张会平,金大卫.文本信息可视化模型研究[J].情报学报,2007,26(1):155-160.

[9] 杨 峰,周 宁,吴佳鑫.基于信息可视化技术的文本聚类方法研究[J].情报学报,2005,24(6):679-683.

[10] 张兆锋,张志平,乔晓东,等.信息可视化在科技文献深度挖掘中的应用[J].情报学报,2007,26(3):408-414.

[11] 陈 颖,张学福,姜世华等.文档类型信息检索可视化系统比较分析[J].情报杂志,2010,29(1):169-172.

[12] 安海忠,崔 娜.专题新闻文本集信息可视化研究[J].图书情报工作,2009,53(12):117-120.

[13] 刘 永,王素立.竞争情报可视化探讨[J].图书馆论坛,2009,29(6):158-160.

[14] 申 莹.针对确定话题的离散文本舆情聚类与可视化研究[D].武汉:武汉理工大学,2011.

[15] 郭建忠,成 毅,傅文棋,等.突发事件网络舆情可视化研究与实现[J].地矿测绘,2012,28(4):5-8.

[16] 许 星,席鹏富,秦 天,等.社会网络的舆情信息分析与可视化——以新浪微博为例[J].计算机光盘软件与应用,2013,(12):94-95.

[17] Havre S, Hetzler E, Whitney P, et al. ThemeRiver: Visualizing thematic changes in large document collections[J]. In IEEE Transactions on Visualization and Computer Graphics, 2002,8(1): 9-20.

[18] 高承实,荣 星,陈 越,等.微博舆情监测指标体系研究[J].情报杂志,2011,30(9):66-70.

[19] 曾润喜,徐晓林.网络舆情突发事件预警系统、指标与机制[J].情报杂志,2009,28(11):52-54.

[20] Bing Liu. Sentiment Analysis and Opinion Mining[M]. Morgan & Claypool Publishers, 2012: 3-5.

[21] Jeffrey Heer, Stuart K Card, James A Landay. Prefuse: A toolkit for interactive information visualization[C]. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05). ACM, New York, NY, USA, 2005:421-430.

[22] 4·11兰州自来水苯含量超标事件[EB/OL].[2014-06-15]. http://baike.sogou.com/v67606675.htm.

[作者简介]詹 建，男，1973年生，兰州大学信息科学与工程学院讲师。

高民权，男，1991年生，兰州大学信息科学与工程学院本科生(已毕业)。

收稿日期:2014-07-23